# Contrôle Science des données 2

L3 MIASHS

22 octobre 2023

1h30

Notes de cours et TP autorisées, documentation concernant python et ses bibliothèques autorisée.

Tout recours à des outils tels que (mais pas seulement) ceux connus sous le nom d’ « IA génératives » est évidemment strictement interdit (et totalement inutile).

Prénom :

Nom :

Ci-dessous vous indiquez la réponse aux questions. Toutes les commandes python doivent être fournies dans un fichier Prénom-Nom.cc-sd2.py avec le compte-rendu en **pdf**.

Rappel : le fichier python que vous m’envoyer doit s’exécuter sans erreur.

Remarque : le style des paragraphes où vous indiquez les réponses aux questions se nomme réponse. Merci d’utiliser ce style : il affiche le texte en bleu ; si votre réponse n’apparaît pas en bleu, c’est que vous n’utilisez pas le bon style.

**Conseil : réfléchissez avant d’agir !**

# Première partie : question de cours

Le sur-apprentissage consiste à utiliser plus de données qu’il n’en faut pour construire un modèle. Il n’a pas de conséquence sur la capacité de prédiction du modèle.

Indiquez clairement votre réponse ci-dessous :

* Vrai
* Faux
* Je ne sais pas

Plus un modèle est « gros », meilleures sont ses capacités de prédiction.

Indiquez clairement votre réponse ci-dessous :

* Vrai
* Faux
* Je ne sais pas

Un jeu d’exemples est constitué de 300 individus répartis entre 4 classes. Il y a 120, 80, 60 et 40 exemples dans chacune. Quelle est l’entropie de cette population ? Indiquez le calcul et le résultat.

# Deuxième partie

On utilise le jeu de données disponible à l’url

<https://philippe-preux.github.io/ensg/miashs/l3-sd2/tp/cc1/flea.csv>.

Ce jeu de données concerne des puces. La première colonne (species) est un attribut catégorique qui indique la classe de l’individu. Les autres colonnes indiquent des mesures de différentes caractéristiques d’un individu.

Vous répondez aux questions suivantes. Votre fichier Python doit contenir les commandes permettant de répondre aux questions. Indiquez clairement à quelle question vous répondez dans le fichier Python.

1. Dans le fichier Python, chargez le fichier dans un objet dénommé flea.
2. Nombre de données (individus) contenues dans ce jeu de données ?
3. Quel est le nombre d'attributs décrivant les données ?
4. Quelle est la liste des attributs ?

Ce jeu de données contient des exemples de 3 classes différentes. La classe est spécifiée dans l’attribut species.

1. Combien y a-t-il d’exemples de chaque classe ?
2. On réalise une exploration visuelle du jeu de données en vue de déterminer si la prédiction de la classe semble faisable par un arbre de décision. Ci-dessous, vous expliquez comment vous procédez et ce que vous constatez ; n’hésitez pas à insérer des figures qui vous paraîtraient utiles pour justifier votre démarche. Un arbre de décision semble-t-il pouvoir déterminer parfaitement la classe (species) ? Dans votre fichier Python, vous indiquez les instructions que vous exécutez afin de répondre à cette question. N’hésitez pas à y mettre des commentaires pour que l’on comprenne ce que vous faites.

On veut induire un arbre de décision pour prédire la classe (species). Indiquez les instructions que vous effectuez dans le fichier Python.

1. Insérez ci-dessous une représentation graphique de l’arbre obtenu. Faites-en sorte qu’il soit joli et lisible.
2. Quelle est l’erreur de test de cet arbre ? Vous mesurez cette erreur en induisant 100 arbres de décision dont vous mesurez le taux de succès et vous en indiquez la moyenne et l’écart-type.
3. Quelle est l’erreur de test de cet arbre pour chacune des classes ?

# Troisième partie

On utilise le jeu de données disponible à l’url

[https://philippe-preux.github.io/ensg/miashs/l3-sd2/tp/cc1/australian-crabs.csv.](https://philippe-preux.github.io/ensg/miashs/l3-sd2/tp/cc1/australian-crabs.csv)

Ce jeu de données concerne des crabes. Les deux premières colonnes sont des attributs catégoriques, une couleur indiquant l’espère et le genre : chacun est une classe. Le troisième attribut est juste un index de l’exemple dans le jeu de données ; il n’a aucun sens. Les autres attributs indiquent les mesures de différentes caractéristiques d’un individu.

Vous répondez aux questions suivantes. Votre fichier Python doit contenir les commandes permettant de répondre aux questions. Indiquez clairement à quelle question vous répondez dans le fichier Python.

1. Dans le fichier Python, chargez le fichier dans un objet dénommé crabs.
2. Nombre de données contenues dans ce jeu de données ?
3. Quel est le nombre d'attributs décrivant les données ?
4. Quelle est la liste des attributs ?
5. On veut prédire l’attribut Species. Réaliser une exploration visuelle du jeu de données en vue de déterminer si la prédiction de la classe semble faisable par un arbre de décision. Ci-dessous, vous expliquez comment vous procédez et ce que vous constatez ; n’hésitez pas à insérer des figures qui vous paraîtraient utiles pour justifier votre démarche. Un arbre de décision semble-t-il pouvoir déterminer parfaitement la classe (Species) ? Dans votre fichier Python, vous indiquez les instructions que vous exécutez afin de répondre à cette question. N’hésitez pas à y mettre des commentaires pour que l’on comprenne ce que vous faites.
6. En vous appuyant sur votre réponse à la question précédente, comment faites-vous pour construire un bon (le meilleur possible) arbre de décision ? Vous expliquez ci-dessous votre démarche ; comme dans la deuxième partie de ce contrôle, vous estimez le taux de succès en induisant 100 arbres de décision dont vous mesurez le taux de succès et vous en indiquez la moyenne et l’écart-type. Vous insérez aussi une image de l’un de ces arbres. Dans le fichier python, vous mettez le code qui permet de faire out cela.
7. Pour terminer, on veut maintenant prédire l’attribut Sex. Comment faites-vous ? Quel est le taux de succès ? À nouveau, expliquez votre démarche.